Advanced Engineering Forum *Online: 2012-09-26*
ISSN: 2234-991X, Vols. 6-7, pp 350-355
doi:10.4028/www.scientific.net/AEF.6-7.350

Realization of Bad Message Filtering System Based on kNN

Du Weifeng[a], Min Xiao[b]

School of Mathematics, Physics and Information Engineering, Jiaxing University, Jiaxing 314001, China

[a]woodmud@tom.com, [b]mxggppt@126.com

Keywords: Message Filtration; kNN; Vector Space Model.

Abstract: The transformation is implemented from short message to feature vector in this paper based on the ICTCLAS system. Then we use the kNN method to come to carry on the classified recognition to the message content, thus realizing to filter bad message effectually. The method has been testified effective in experiment.

Introduction

At present, problems caused by the bad short message have become increasingly prominent. Thus our time and energy is wasted greatly and people's normal life is affected. It has caused greater harm to the society and caused the wide attention of whole society. The design of effective short message filtering method can control the bad short message in the proliferation and spread, so as to improve people's working efficiency. But the traditional filtering methods have some limitations, content based filtering is one of the main technologies to solve the problem of bad message. This paper studies the content of some bad short message text structure features as well as the solution. In application of the segmentation system developed by ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), sentence segmentation of message to the specific word conversion is realized, thus laid the foundation for the classification. In this paper, bad message content filtering based on kNN can effectively reduce the bad message disturbance, easy to filter out pornographic information, comment spam and other undesirable content.

The Traditional Filtering Technology

Black List Technology. In the mobile phone bad short message filtering system, a blacklist filtration technology is often used, the process flow as follows[1]**:**

(1) Getting the mobile phone number for sending SMS;
(2) Looking for the mobile phone number in the black list, if it is found, automatically filter;
(3) otherwise, getting SMS text content to notify the user reading;
(4) if be found to be spam, add the number to the blacklist artificially.

Using the black list technology, mobile phone spam messages can be filtered preliminary. When the spam sent by the mobile phone on the blacklist, is filtered directly, but when the spam not sent by the mobile phone on the blacklist, cannot be filtered directly, users must be notified of reading.

Key Words Filtering Technology. When some specific words appeared in the text, then the text is filtered as bad information. But this method has its deficiencies. For example, in the application of filtering the pornographic information, through the detection of the word "sex", it may incorrectly filter the information propagating the anti-pornography.

Theoretical Basis

Vector Space Model. Vector Space Model was proposed by Salton[2] etc. in sixty last century. Vector model changed the limitation of binary weight in Boolean model, and proposed a framework suitable for partial match. Because the vector space model is established in the standard mathematical basis, so the model is most wide in the information retrieval. Vector space model represents the text message with feature and its corresponding weights, so the premise of

application is that the meaning of the text could be reflected through the lexical information. In the process of recognition, the degree of correlation between the text and the query request is described through vector operation.

TFIDF. It is usually with higher accuracy through using a weight value to replace whether the word appears in words weighted system. In order to deal with professional words and common words, they are both the high frequent words, TFIDF (Term Frequency Inverse Document Frequency) is used to measure the weight of a word in text mining areas. The equation is as follows:

$$\mathrm{TFIDF}(W_i, D_j) = \mathrm{TF}(W_i, D_j) \times \mathrm{IDF}(W_i) \ \ (\mathrm{IDF}(W_i) = \log \frac{|D|}{\mathrm{DF}(W_i)}) \tag{1}$$

Wherein W_i represents a particular word, D_j represents the text where the word is in, $\mathrm{TF}(W_i, D_j)$ represents the frequency of the word W_i occurrence in the text D_j, $|D|$ represents the total amount of texts containing in the training set and $\mathrm{DF}(W_i)$ represents he number of texts including the word W_i. To equation (1), we have the following intuitive interpretation:

(1) the word appears more in one text, then the word is more representative to the text, so its weight is greater;

(2) the word appears in more texts, then the word is lower in distinguishing the text categorization, so its weight is smaller.

kNN. The thought of kNN method is: given new text, considered the k texts in the training text which are closest(most similar) to the new text. Then according to the categories of the k texts, the category of new text is determined. The kNN method is a lazy learning method, it has no training phase, its study was delayed to the classification stage, that is to say, passive learning is conducting only when the classification is needed. In this method, the value of k is given, different k values may affect the classification of the final results, as shown in Fig. 1:

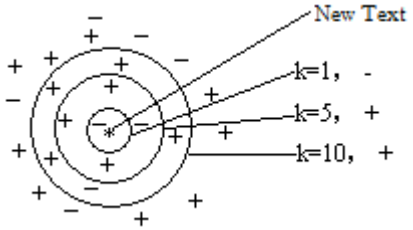

Fig 1 kNN

The Vector Space Model of the Text

Word Segmentation. In order to extract the keywords, the Chinese text should be segmentation processed firstly. Chinese texts use character as the unit, but the basic information bearing unit is word. The specificity of the word is much stronger than the character composing the word. For example: the meaning of the word "Xue Sheng"(student) is single, but the two characters "Xue" and " Sheng " is very rich. From classical Chinese to the vernacular, character set is reduced greatly in Chinese language, but with the development of science and technology, a variety of specialized vocabulary emerge in an endless stream, therefore the word set presents a trend of expansion. This inspires us that text should be decomposed to the word as the unit, rather than the character.

Word segmentation is the process of recognition of word boundary by computer automaticly. From the aspect of form, the input of word segmentation system is the string "C1C2⋯Cn", the output is "W1W2⋯Wm", here, Wi could be monosyllabic word, and also be a plurality of words.

Of course, the word segmentation is not the research content of this paper. It is based on the realization of the Chinese lexical analysis system ICTCLAS by the Institute of computing technology, Chinese Academy of Sciences. The functions of the system are: Chinese word segmentation, POS tagging, named entity recognition, unknown word recognition and etc.. Its segmentation accuracy is up to 97.58%[3].

Stop Words Removal. Stop words usually refer to the high-frequency words like pronouns, prepositions, conjunctions, which are frequent in various types of documents and thought to contribute a few information to the classification. The objective of stop words removal is to remove stop words from the text feature [4].

What is a stop word and what is not a stop word, it has not yet formed a consensus of opinion in the field of text categorization. According to the different specific areas and different preferences that relates to the text classification, the stop words list can be different[4].

After segmentation, and then removal of punctuation, single words and stop words, indexing can be performed on the text. Finally, the text is transformed to the form of word frequency vector.

Indexing of Text. Because the size of the text is not very uniform in the corpus, if absolute frequency of a word used as a measure of important degree standard, will tend to choose big text word as a feature, it is not reasonable. Therefore, text indexing should be performed after the acquisition of the absolute frequency of the word. Text indexing is the process to determine the frequency of various words in the document, and correspond a document to a non-negative real-valued vector --index vector in some way. A relative word frequency as a measure of an important indicator is put forward in this paper. Relative word frequency may have the following two kinds of calculation methods:

(1) the ratio of absolute frequency and the total number of words;

(2) the ratio of absolute frequency and the number of surplus words after filtering treatment.

An algorithm of text indexing is given in this paper. In order to speed up the processing, the index table uses chain address hash list as a storage structure. With a word in a document as hash key, the element of hash table is the following structure:

```
struct
{
    char *key; // the word in the text
    float TF; // index value of the word
}
```

1. Algorithm Description

Algorithm 1(Text indexing algorithm)[6]

Input：The original text

Output：The index vector of the Text

Step 1：Performing segmentation to the original text, calculating the total number n_0 of words

Step 2：Performing filtering based on the segmented text

Step 3：According to the number of words in the text n_1 after step 2, the number of different words n_2 is estimated, then the hash table is built, and begin scanning the text.

① Whether the text scanning is completed, is to step 4; otherwise, the next step.

② The current word is searched in the hash table, if found, then TF++ of the corresponding word; otherwise, put the word in the hash table, TF=1.

③ Scan the next word, turn to①.

Step 4： TF$/n_0$ of the words in the hash table is the index value, the hash table is the index vector the text.

2. Time Complexity Analysis

The time complexity of the algorithm is mainly in step 3. The algorithm is performed only once for scanning the text, process of each word is searching for the hash table. Assuming that text word number is $n_1(n_1 < n_0)$ after step 2, processing time on a text is $1.5n_1$. The algorithm do the same work for each text. If text number is n, the time complexity of the training set is $O(n \times n_1)$.

3. Space Complexity Analysis

The space complexity of the algorithm is mainly reflected in the hash table storage space. It is equal to the count $n_2(n_2 < n_1)$ of different words in the text in count. The algorithm do the same work for each text. If text number is n, the space complexity of the training set is $O(n \times n_2)$.

Extraction of Text Feature

The dimension of words in a text is very high, may be up to several tens of thousands (Modern Chinese characters are commonly used less, just five thousand or six thousand, but words are commonly used many, there are tens of thousands of different words), but they have different contribution. In order to reduce the computational complexity, we need to measure the weight of words in the text. Only the words which are larger than a certain threshold weight can be as keywords of text content. Usually we select keywords instead of words collection occurring in the text. Extraction of keywords is also known as the extraction of text feature. Extraction of feature can reduce overfitting. A reasonable and effective method to measure the word weight is the important technical basis of extraction of text feature.

The essence of text feature extraction is dimensionality reduction technique of high dimensional data, or to say, the high dimensional data maps to a lower dimensional space by transformation. The main problem of dimension reduction methods is that information contains in data may lose from the high dimensionality to low dimensional. So the original different data in a high dimensional space will be mixed together in low-dimensional space(Fig. 2). Therefore, finding a right mapping is the key to transform from a higher dimensional space to a lower dimensional space[5].

Fig. 2 High Dimensional Data into a Low Dimensional Space

Using TFIDF as the measure index can be more appropriately express the fact. For larger TF value and smaller IDF value, the word in almost all texts have higher frequency, commom words are so. If the TF value and IDF value are larger, the words are very important. They appear in some texts with high frequency and only appear in a few texts. Such words can represent the keywords of the text, they are the the professional words representing the content of the text generally.

In general, short words are function-oriented with higher frequency and more meaning. And long words are content-oriented with lower frequency. Long words have a higher degree of specificity, so the weight should be increased.

Of course, there are many kinds of methods of text feature extraction. We can use different evaluation functions such as information gain, expected cross entropy, textual evidence weight.

By formula (1), TFIDF value of all words in the text can be calculated. All the words in a text sorted by the TFIDF value, each text will correspond to a sorted list. The sorted lists of the texts in the same category will be similar.

After the above treatment, the list of words have a corresponding real numbers. They can be sorted from large to small. At this point, feature extraction is relatively simple, according to need can be in one of two ways:

(1) choosing a fixed number of keywords with largest weight;

(2) Selecting keywords the weight of which is greater than a threshold.

In fact, the words with smaller weight have not too much significance for text classification. They are some general words or some uncommon words. We can improve the generalization ability of knowledge and enhance the knowledge of fitness after dimension reduction treatment. According to the experimental results, the two methods have their own advantages and disadvantages, the first

way can guarantee the key coverage, but sometimes we can't choose the most suitable number of keywords, because different text subject to different concepts, themes of the dispersion is also different; the second way to choose the subject and content of the relationship between relative close, but for the theme of the text, choose the subject may be either too little or too much.

Basic Steps of KNN

Step 1: Collecting a certain amount of normal messages and bad messages,2/3 of them as training set,1/3 as testing set.

Step 2: Processing the training set, then getting knowledge:

① Performing word segmentation to the messages in training set with ICTCLAS, then removing stop words, single words etc., getting the words and the corresponding relative word frequency TF value;

② According to the method of feature vector selection, extracting, 1000(adjustable) words as keywords;

③ Keywords and the corresponding TF value forming feature vectors, a training set of all message constituting a vector space, forming the knowledge base of k nearest neighbor method.

Step 3：Reading the unknown category message, extracting the keywords and the corresponding relative word frequency TF values form the text feature vector using the method in step 2;

Step 4：Calculating the cosine value of vector between the message and the vectors in knowledge base. According to the category of the largest k texts, determining the message category by vote.

Analysis and Evaluation of the Results

For the short message filtering system based on the kNN designed in the paper, the recall and precision is listed in table 1:

Table 1 Accuracy

Method	Macro average recall	Macro average precision	Micro average
kNN	80.27%	82.38%	81.80%

Because the text length of SMS is generally around 50 words, so the vector dimension is small, for each text less information can be gotten, so the accuracy is about 80%, but it has a certain value of application.

Conclusions

The application of kNN classification in short message filtering system is discussed briefly in this paper. The theoretical basis and the key algorithm is given. The results kNN classification model of are analyzed and assessed. It is confirmed that the method to short message filter has a certain practicality.

Acknowledgements

This work was partially supported by Zhejiang province fatal project (priority subjects) key industrial project (Grant No: 2008C11011), the National Nature Science Foundation of China (Grant No: 61070213, 60875034), the Specialized Research Fund for the Doctoral Program of Higher Education of China(No.20060613007) and the Provincial Nature Science Foundation of Zhejiang (Grant No: LY12A01019, LY12F02019).

References

[1] Li Lu, Qin Weiping. Superficial Analysis about the Use of Bayes Classifies Method in Filtration System of Trash Short Message. Technology Square, 2007.7

[2] G. Salton, A. Wong, C.Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11):613~620, 1975

[3] http://www.nlp.org.cn/project/project.php?proj_id=6

[4] Song Fengxi. Research on several basic problems of automatic text classification. Doctoral Dissertation of Nanjing University of Science and Technology. 2004

[5] Zhang Yuntao, Gong Ling. Data mining principle and technology. Beijing: Publishing House of electronics industry，2004

[6] Du Weifeng. Application of Rough Set Theory in Chinese Text Categorization. Doctoral dissertation of Southwest Jiao Tong University. 2006